(19)日本国特許庁（ＪＰ）　(12)公開特許公報（Ａ）　(11)特許出願公開番号

特開平9－319768

(43)公開日　平成９年(1997)12月12日

(51)Int.Cl.[6]		識別記号	庁内整理番号	ＦＩ			技術表示箇所
Ｇ０６Ｆ	17/30			Ｇ０６Ｆ	15/401	３２０Ａ	
	15/18	５６０			15/18	５６０Ｊ	

審査請求　未請求　請求項の数２　ＦＤ　（全 ５ 頁）

(21)出願番号　特願平8－157723

(22)出願日　平成８年(1996)５月29日

(71)出願人　000000295
沖電気工業株式会社
東京都港区虎ノ門１丁目７番12号

(72)発明者　福本　淳一
東京都港区虎ノ門１丁目７番12号　沖電気工業株式会社内

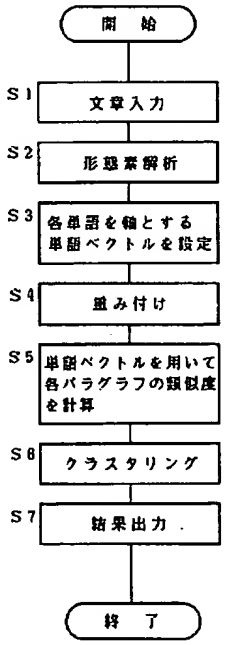
(74)代理人　弁理士　佐藤　幸男　（外１名）

(54)【発明の名称】　要点抽出方法

(57)【要約】

【課題】　文章中の各文中の語句の参照関係や文章の構造情報等を用いることなく、文章中の重要部分を抽出する。

【解決手段】　要点抽出対象の文章が入力されると（ステップＳ１）、形態素解析を行って（ステップＳ２）、単語を抽出する。次いで、各単語を軸とする単語ベクトルを設定すると共に（ステップＳ３）、各単語に重み付けの値を付与する（ステップＳ４）。更に、各パラグラフの類似度を単語ベクトルの値を用いて計算する（ステップＳ５）。このパラグラフの類似度からクラスタを生成し（ステップＳ６）、その結果を出力する（ステップＳ７）。



本発明方法のフローチャート

1

【特許請求の範囲】

【請求項１】　自然言語で記述された文章中の各パラグラフを、当該文章中に出現する各単語を軸とし、各軸に対して、各々のパラグラフに出現する回数を対応させた単語ベクトルで表し、

各パラグラフの単語ベクトルの類似度を計算し、

前記単語ベクトルの最も類似度の高いパラグラフを、前記文章における最も重要な部分として出力することを特徴とする要点抽出方法。

【請求項２】　請求項１記載の要点抽出方法において、任意の単語は、予め決められた基準により重み付けされることを特徴とする要点抽出方法。

【発明の詳細な説明】

【０００１】

【発明の属する技術分野】本発明は、自然言語で記述された文章から重要部分を抽出する要点抽出方法に関するものである。

【０００２】

【従来の技術】従来の自然言語で記述された文章から重要部分を抽出する方法においては、文章中の各文の形態素解析、構文解析を行い、各文中の語句の参照関係に基づき、他の文から最も多く参照されている文を重要文としたり、接続語句等を用いた文章の構造情報を用いて重要部分を判定するといったヒューリスティックスに基づく手法が用いられていた。

【０００３】

【発明が解決しようとする課題】しかしながら、一般に、文章中の各文中の語句の参照関係の解析のためには、その文章中で用いられる語句に関する上位－下位の知識等の多くの知識が必要であり、参照関係の解析そのものが困難であるといった問題があった。また、接続語句等を用いて得られた文章の構造情報から重要部分を判定する方法もあるが、そのためには文章の構造を解析するための接続語句等の情報を多く登録しておかなければならないといった問題があった。

【０００４】このような点から、文章中の各文中の語句の参照関係や文章の構造情報等を用いることなく、文章中の重要部分を抽出することのできる要点抽出方法の実現が望まれていた。

【０００５】

【課題を解決するための手段】本発明は、前述の課題を解決するため次の構成を採用する。

〈請求項１の構成〉自然言語で記述された文章中の各パラグラフを、その文章中に出現する各単語を軸とし、各軸に対して、各々のパラグラフに出現する回数を対応させた単語ベクトルで表し、各パラグラフの単語ベクトルの類似度を計算し、単語ベクトルの最も類似度の高いパラグラフを、文章における最も重要な部分として出力することを特徴とする要点抽出方法である。

【０００６】〈請求項１の説明〉請求項１の発明は、文章中、重要な部分は、繰り返し述べられていることが多いという点に着目し、同様な事柄が述べられているパラグラフをその文章の要点であると判定するようにしたものである。ここで、パラグラフとしては、文章中の章単位、文単位等、任意の区切りであってもよい。また、軸とする単語は、文章中に出現する全ての単語を対象とするが、適宜選択するようにしてもよい。

2

【０００７】このように、文章中の単語のみの情報を用いて、それらの情報の統計的処理により、文章中の重要なパラグラフを抽出するようにしているため、文章中の各文中の語句の参照関係や文章の構造情報等を用いることなく、容易かつ正確に重要部分を抽出することが可能となる。

【０００８】〈請求項２の構成〉請求項１記載の要点抽出方法において、任意の単語は、予め決められた基準により重み付けされることを特徴とする要点抽出方法である。

【０００９】〈請求項２の説明〉予め決められた基準とは、例えば単語の品詞情報に基づく基準である。即ち、助詞、助動詞等の付属語は文章中に多く出現するが、これらの付属語情報は文章の重要度の判定には必要ないため、低い重み付けの値を設定する。一方、文章中の各単語のうち、名詞や動詞等の自立語は文章中で重要部分を判定するために必要であるため、高い重み付けの値を設定する。これにより、各パラグラフに対して単語の重要度を考慮した単語ベクトルが設定される。また、この重み付けの基準は、単語の品詞情報だけでなく、これ以外にも、ユーザによって特定の単語を指定するといったように、適宜選択が可能である。

【００１０】

【発明の実施の形態】以下、本発明の実施の形態を図面を用いて詳細に説明する。図１は本発明の要点抽出方法を示すフローチャートであるが、この説明に先立ち、本発明の要点抽出方法を実現するための要点抽出装置を説明する。

【００１１】図２は、その要点抽出装置を示す構成図である。図の装置は、入力部１、形態素解析処理部２、重要部分抽出部３、出力部４、重み付け処理部５からなる。

【００１２】要点抽出装置は、マイクロコンピュータで構成され、入力部１は、例えば入力インタフェースやキーボードといった解析対象文の入力を行う部分である。また、形態素解析処理部２は、入力部１に入力された文を各単語に分割する処理を行う機能を有している。

【００１３】重要部分抽出部３は、形態素解析された単語情報と各単語に付与された重み付け情報を用いて重要部分であるパラグラフを抽出する機能を有している。即ち、この重要部分抽出部３は、重要な部分は、文章中で繰り返し出現することが多いという点に着目し、同様な事柄が述べられているパラグラフ、つまり、共通してい

る単語が最も多いパラグラフを、その文章における要点が記述されているパラグラフとして出力するようにしたものである。

【００１４】重み付け処理部５は、形態素解析された単語情報に対して、重み付けの計算を行う機能を有している。また、出力部４は、例えば、表示装置や印刷装置といった出力部であり、重要部分抽出部３で抽出された重要部分の出力を行う機能を有している。尚、上記の形態素解析処理部２～重み付け処理部５は、各機能を実現するプログラムと、これを実行するマイクロコンピュータにおける中央処理装置やメモリといった制御部により構成されている。

【００１５】次にこのように構成された要点抽出装置を用いた要点抽出方法を図１に沿って説明する。先ず、ユーザは、要点抽出を行う文を入力部１に入力する（ステップＳ１）。これにより、形態素解析処理部２は文章中の各パラグラフにおける単語の認識処理を行う（ステップＳ２）。尚、この形態素解析処理については既知の処理であるため、ここでの説明は省略する。

【００１６】次に、重要部分抽出部３は、文章中の各パラグラフ中に存在する全ての単語情報に対して、各単語を軸とする単語ベクトルを設定する（ステップＳ３）。図３は、各パラグラフの単語ベクトルの説明図である。この例は、三つの単語で四つのパラグラフの場合を示しており、図中、軸６、７、８がそれぞれ、単語１、単語２、単語３を示し、９～１２が、各パラグラフに対する単語ベクトルを示している。

【００１７】また、重要部分の抽出のためには、単語の重要度が異なるため、重み付け処理部５は、それを表す重み付けの値を、抽出された各単語に対して付与する（図１におけるステップＳ４）。この重み付けの値を与える方法としては、例えば、文章中の各単語の品詞情報を用いる方法がある。これは、文章中の各単語のうち、助詞、助動詞等の付属語は文章中に多く出現するが、これらの付属語情報は文章の重要度の判定には必要ないため、低い重み付けの値を設定する。一方、文章中の各単語のうち、名詞や動詞等の自立語は文章中で重要部分を判定するために必要であるため、高い重み付けの値を設定する。そして、文章中の各単語の品詞情報から設定された重み付けの値を文章中の各パラグラフの単語ベクトルに対して掛け合わせる。これにより、各パラグラフに対して単語の重要度を考慮した単語ベクトルが設定される。尚、このような重み付けの値の付与は、これ以外にも、ユーザが、重み付け処理部５に対して特定の単語を任意の値を指定できるよう構成してもよい。

【００１８】次に、文章中の各パラグラフに対して、類似度の計算を、上記の重み付けされた単語ベクトルを用いて行う（ステップＳ５）。即ち、単語ベクトルの値が類似しているパラグラフを類似度の高いパラグラフであると判断する。尚、このような類似度の計算方法として、例えば「G. Salton：Automatic Text Processing, Addison－Wesley Publishing Company (1989)」Chapter 10等に示されている方法を用いることができる。

【００１９】全てのパラグラフの類似度の計算が済むと、その類似度を用いて各パラグラフについてクラスタリングを行う（ステップＳ６）。そして、クラスタリングが終了すると、その結果を、出力部４に出力し（ステップＳ７）、要点抽出処理を終了する。

【００２０】次に、上記の動作を更に詳細に説明する。文章中の各パラグラフに対して、重み付けされた単語ベクトルに基づきその類似度の計算による重要部分の抽出の手法を以下に示す。

【００２１】図４は、要点抽出処理における演算式の説明図である。パラグラフの類似度計算の方法としては、パラグラフの単語ベクトルを$X=(x_1, x_2, \cdots, x_t)$，$Y=(y_1, y_2, \cdots, y_t)$とした場合、単語ベクトルX，Yの類似度は、図中の式（１）で表される。

【００２２】また、単語１、単語２の単語ベクトルをそれぞれ$W1=(w1_1, w1_2, \cdots, w1_t)$，$W2=(w2_1, w2_2, \cdots, w2_t)$とし、単語の重み付けベクトルを$A=(a_1, a_2, \cdots, a_t)$としたとき、単語ベクトルW１，W２の類似度$S_{12}$は、図中の式（２）で計算される。尚、このとき、ｔは単語の種類数である。

【００２３】以上の類似度の計算を文章中の全てのパラグラフ（１，…，ｎ）について計算した結果は、図中の配列（３）のように示される。尚、ここで、S_{ij}はパラグラフｉとパラグラフｊの類似度を計算した値であるとする。但し、$S_{ii}=0$であるとする。例えば、配列（４）は、パラグラフ１，２，３，４からなる文章について得られた配列を示している。

【００２４】次に、以上のようにして得られた各パラグラフ間の類似度の値を用いてパラグラフのクラスタリングを行う。このクラスタリングの方法としては、上述した「G. Salton：Automatic Text Processing, Addison－Wesley Publishing Company (1989)」Chapter 10において示されている方法を用いることが可能である。

【００２５】図５は、クラスタリングの一例を示す図である。この例は、上記図４の配列（４）で示したパラグラフのクラスタリングを示している。即ち、配列（４）において、類似度の最も高いのは、S_{14}およびS_{41}の０．９である。従って、文章中、パラグラフ１とパラグラフ４とが最も類似度の高いパラグラフであるため、これらのパラグラフからクラスタを生成する。

【００２６】次に、パラグラフ１またはパラグラフ４のどちらかのパラグラフと類似度の高いパラグラフを抽出する。ここで、パラグラフ１と最も類似度の高いパラグラフの値は、パラグラフ２との０．７であり、また、パ

5

ラグラフ４と最も類似度の高い値は、パラグラフ２との０．５である。従って、パラグラフ１とパラグラフ４とで生成したクラスタとパラグラフ２とで上位のクラスタを生成する。

【００２７】以下、同様に、生成したクラスタのうち、いずれかのパラグラフと最も類似度の高いパラグラフで、更に上位のクラスタを生成する。ここでは、パラグラフが４個であるため、残りのパラグラフ３によって上位のクラスタが生成される。尚、クラスタの生成を、いずれかのパラグラフの一方との比較ではなく、二つのパラグラフの合成ベクトルとの比較によって行うようにしてもよい。

【００２８】そして、このようなクラスタリングの結果が出力部４から出力される。これにより、ユーザは、パラグラフ１とパラグラフ４とが最も重要なパラグラフであることを知ることができる。即ち、文章中、パラグラフ１とパラグラフ４とで同様の事柄が最も多く述べられているため、これらのパラグラフで文章の要点が記されていることが分かる。

【００２９】以上のように、上記具体例では、文章中の各パラグラフの単語のみの情報を用い、それらの情報の統計的処理により、パラグラフ間の類似度を計算して、

6

文章中の重要部分の判定を行うようにしたので、文章中の各文中の語句の参照関係や文章の構造情報等を用いることなく、容易にかつ正確に要点を抽出することができる。

【図面の簡単な説明】

【図１】本発明の要点抽出方法を示すフローチャートである。

【図２】本発明の要点抽出方法を実現するための要点抽出装置の構成図である。

【図３】本発明の要点抽出方法における各パラグラフと単語ベクトルの説明図である。

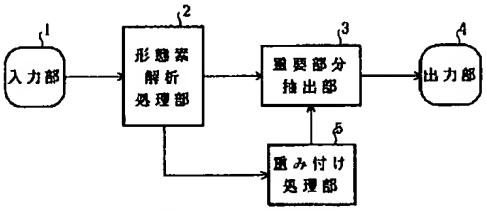
【図４】本発明の要点抽出方法における演算式の説明図である。

【図５】本発明の要点抽出方法におけるクラスタリングの一例を示す図である。

【符号の説明】

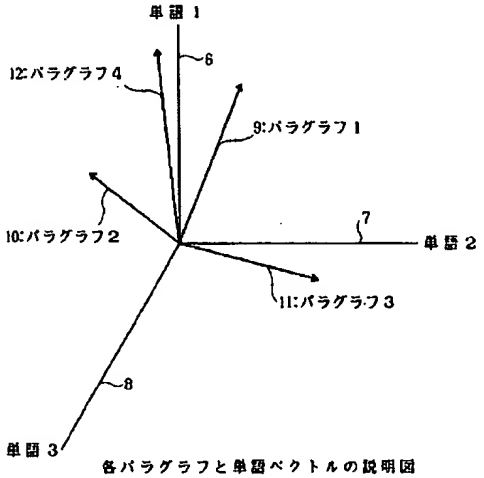
1　入力部
2　形態素解析処理部
3　重要部分抽出部
4　出力部
5　重み付け処理部

【図２】



要点抽出装置の構成図

【図３】



各パラグラフと単語ベクトルの説明図

【図４】

(1)　$\sum_{i=1}^{t} x_i \cdot y_i$

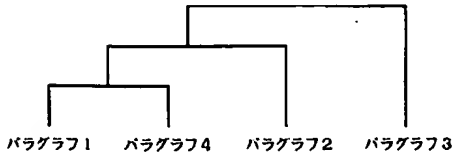
(2)　$S_{12}=\sum_{i=1}^{t} a_i w1_i \cdot a_i w2_i$

(3)　$\begin{pmatrix} S_{11} & S_{12} & \cdots & S_{1n} \\ S_{21} & S_{22} & \cdots & S_{2n} \\ \cdots & & & \cdots \\ S_{n1} & S_{n2} & \cdots & S_{nn} \end{pmatrix}$

(4)　$\begin{pmatrix} S_{11} & S_{12} & S_{13} & S_{14} \\ S_{21} & S_{22} & S_{23} & S_{24} \\ S_{31} & S_{32} & S_{33} & S_{34} \\ S_{41} & S_{42} & S_{43} & S_{44} \end{pmatrix} = \begin{pmatrix} 0 & 0.7 & 0.6 & 0.9 \\ 0.7 & 0 & 0.5 & 0.5 \\ 0.6 & 0.5 & 0 & 0.4 \\ 0.9 & 0.5 & 0.4 & 0 \end{pmatrix}$

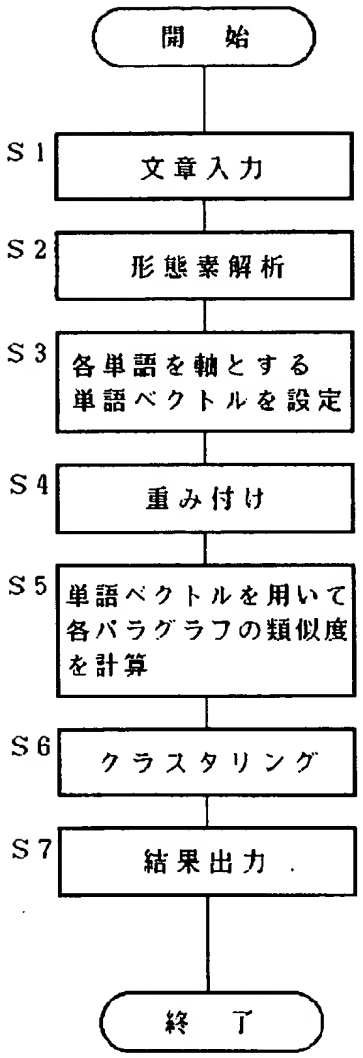
演算式の説明図

【図５】



クラスタリングの一例

【図1】



本発明方法のフローチャート